

# PATENT ABSTRACTS OF JAPAN

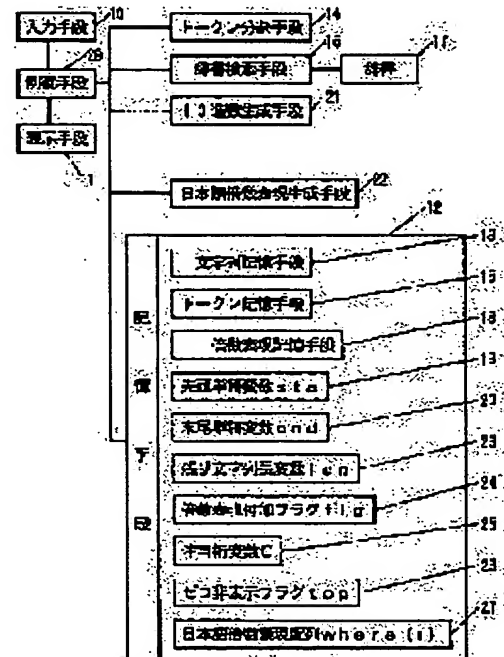
(11) Publication number : 07-319871  
(43) Date of publication of application : 08.12.1995

(51) Int. Cl. G06F 17/27

(21) Application number : 06-108362 (71) Applicant : MATSUSHITA ELECTRIC IND CO LTD  
(22) Date of filing : 23. 05. 1994 (72) Inventor : KINOSHITA HITOMI

## (54) MORPHEME ANALYSIS DEVICE

(57) Abstract:  
PURPOSE: To precisely analyze the expression of an English multiple.  
CONSTITUTION: An input means 10 inputting an English character string, character string storage means 13 storing the English character string inputted from the input means 10, a token division means 14 dividing the English character string stored in the character string storage means 13 for respective tokens and a token storage means 15 storing the tokens divided by the token division means 14 are provided. A multiple expression storage means 18 storing the expression of the English multiple and a decimal number generation means 21 generating a decimal number from the token expressing a numerical value by referring to the tokens stored in the token storage means 15 and the expression of the English multiple stored in the multiple expression storage means 18 are provided.



**LEGAL STATUS**

[Date of request for examination]  
[Date of sending the examiner's decision of rejection]  
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]  
[Date of final disposal for application]  
[Patent number]  
[Date of registration]  
[Number of appeal against examiner's decision of rejection]  
[Date of requesting appeal against examiner's decision of rejection]  
[Date of extinction of right]

Copyright (C) ; 1998, 2003 Japan Patent Office

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号

特開平7-319871

(43)公開日 平成7年(1995)12月8日

(51)Int.Cl.<sup>8</sup>

G 0 6 F 17/27

識別記号

庁内整理番号

F I

技術表示箇所

8219-5L

G 0 6 F 15/ 38

E

審査請求 未請求 請求項の数2 O L (全 7 頁)

(21)出願番号 特願平6-108362

(22)出願日 平成6年(1994)5月23日

(71)出願人 000005821

松下電器産業株式会社

大阪府門真市大字門真1006番地

(72)発明者 木下 ひとみ

大阪府門真市大字門真1006番地 松下電器  
産業株式会社内

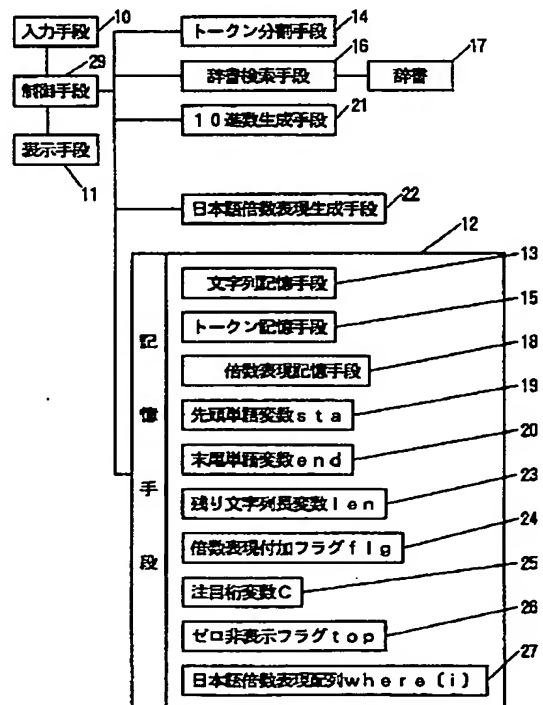
(74)代理人 弁理士 小銀治 明 (外2名)

(54)【発明の名称】 形態素解析装置

(57)【要約】

【目的】 英語の倍数表現を正しく解析する。

【構成】 英文字列を入力する入力手段10と、入力手段10から入力された英文字列を記憶する文字列記憶手段13と、文字列記憶手段13に記憶された英文字列をトークン毎に分割するトークン分割手段14と、トークン分割手段14により分割されたトークンを記憶するトークン記憶手段15とを備え、英語倍数表現を記憶する倍数表現記憶手段18と、トークン記憶手段15に記憶されたトークンと倍数表現記憶手段18に記憶された英語倍数表現とを参照して、数値を表現するトークンから10進数を生成する10進数生成手段21とを有する。



## 【特許請求の範囲】

【請求項 1】 原文文字列を入力する入力手段と、前記入力手段から入力された原文文字列を記憶する文字列記憶手段と、前記文字列記憶手段に記憶された原文文字列をトークン毎に分割するトークン分割手段と、前記トークン分割手段により分割されたトークンを記憶するトークン記憶手段とを備え、

原語倍数表現を記憶する倍数表現記憶手段と、前記トークン記憶手段に記憶されたトークンと前記倍数表現記憶手段に記憶された原語倍数表現とを参照して、数値を表現するトークンから、そのトークンに相当する 10 進数を生成する 10 進数生成手段とを有することを特徴とする形態素解析装置。

【請求項 2】 前記 10 進数生成手段が生成した 10 進数に日本語倍数表現を付加する日本語倍数表現生成手段を有することを特徴とする請求項 1 記載の形態素解析装置。

## 【発明の詳細な説明】

## 【0001】

【産業上の利用分野】 本発明は入力した原文文字列の形態情報を出力する形態素解析装置に関するものである。

## 【0002】

【従来の技術】 形態素解析機能を備えた機械翻訳装置により、原文文字列を翻訳するに先立ち、入力した原文文字列をトークン毎に分割し、各トークンに対応する辞書情報を獲得する形態素解析が行われる。

## 【0003】

【発明が解決しようとする課題】 さて従来の形態素解析装置では、原文を英文として例をあげると、次に述べるように、入力した英文文字列に英語倍数表現が含まれていると誤った形態素情報が出力され、誤訳の原因となるという問題点があった。例えば「two hundred and twenty-two」という英語倍数表現が英文文字列に含まれているとき、従来の形態素解析装置ではスペースを区切り文字として区切り文字に挟まれた英文字をトークンとするので、この英文文字列は、「two」、「hundred」、「and」、「twenty-two」の 4 つのトークンに分割される。そして、このまま機械翻訳装置により翻訳を行うと、「にひやくと にじゅうに」ないし「にひやく と にじゅうに」という訳語が得られる。もちろん、この際、正しい訳は「にひやくにじゅうに(二百二十二、222)」である。ここで、英語では「hundred」、「thousand」、「million」、・・・というような倍数表現の次に「and」が付されることがあり、従来の形態素解析装置ではこのような「and」が英文文字列に含まれていると、この「and」の前後を分割してしまうものであった。また以上、基数詞について述べたが、「first」、「second」、・・・というような序数詞についても同様の問題点がある。

【0004】 そこで本発明は、倍数表現を含む原文文字列を正しく形態素解析することができる形態素解析装置を提供することを目的とする。

## 【0005】

【課題を解決するための手段】 本発明の形態素解析装置は、原文文字列を入力する入力手段と、入力手段から入力された原文文字列を記憶する文字列記憶手段と、文字列記憶手段に記憶された原文文字列をトークン毎に分割するトークン分割手段と、トークン分割手段により分割されたトークンを記憶するトークン記憶手段とを備え、原語倍数表現を記憶する倍数表現記憶手段と、トークン記憶手段に記憶されたトークンと倍数表現記憶手段に記憶された原語倍数表現とを参照して、数値を表現するトークンから、そのトークンに相当する 10 進数を生成する 10 進数生成手段とを有する。

## 【0006】

【作用】 上記構成により、入力された原文文字列に倍数表現が含まれている際、この倍数表現が、倍数表現記憶手段及び 10 進数生成手段により、10 進数に変換される。

## 【0007】

【実施例】 次に図面を参照しながら本発明の実施例を説明する。図 1 は本発明の一実施例における形態素解析装置のブロック図、図 2 は本発明の一実施例における形態素解析装置の機能ブロック図である。

【0008】 尚、本実施例においては、英文を原文として説明する。図 1 において、1 は図 4～図 8 のフローチャートに沿う制御プログラムを記憶する ROM（リードオンリーメモリ）、2 は ROM1 の制御プログラムを実行し他の各要素を制御する CPU（中央処理装置）、3 は後述する各記憶領域などが設けられている RAM（ランダムアクセスメモリ）、4 は原稿を読み取り文字コードに変換して出力する OCR（光学式文字読取装置）、5 はユーザが必要な情報を入力するためのキーボード、6 はユーザに必要な情報を表示するための CRT（カソードレイチューブ）、7 は辞書が設けられたハードディスク装置である。

【0009】 図 2 において、10 は英文字列を入力する入力手段であり、図 1 の OCR 4、キーボード 5 がこれに対応する。11 は処理状況などを表示する表示手段であり、図 1 の CRT 6 がこれに対応する。12 は RAM 3 に設けられる記憶手段であり、13 は入力手段 10 から入力された英文字列を記憶する文字列記憶手段、14 は文字列記憶手段 13 内の英文字列を、区切り文字（スペース、カンマ、セミコロン、コロンの、感嘆符、疑問符、カッコなど）を参照しながらトークンに分割するトークン分割手段、15 は分割されたトークン群を記憶するトークン記憶手段、16 はハードディスク装置 7 に格納された辞書 17 を検索する辞書検索手段、18 は図 3 に示すような英語倍数表現パターン、倍数（詳細は後

## 3

述)などを記憶する倍数表現記憶手段、19は数値を表示するトークンの先頭単語が代入される先頭単語変数  $s$   $t$   $a$ 、20は同末尾単語が代入される末尾単語変数  $e$   $n$   $d$ 、21はトークン記憶手段15に記憶されたトークンと倍数表現記憶手段18に記憶された原語倍数表現とを参照して数値を表示するトークンから10進数を生成する10進数生成手段、22は10進数生成手段21が生成した10進数に「億」、「万」等の日本語倍数表現を付加する日本語倍数表現生成手段である。また23は日本語倍数表現生成手段22が駆動される際、10進数のうち未だ処理をされていない残り文字列の長さを示す残り文字列変数  $l$   $e$   $n$ 、24は日本語の倍数表現を付加すべきとき「1」、そうでないとき「0」の値を持つ倍数表現付加フラグ  $f$   $l$   $g$ 、25は10進数のうち現在注目している注目桁を示す注目桁変数  $C$ 、26は「億」などの日本語の倍数表現に後続する0を表示すべきでないとき「1」、表示すべきとき「0」の値を持つゼロ非表示フラグ  $t$   $o$   $p$  である。ここで、日本語の倍数表現に続く0は通常表示しない(例えば、「1億0030万」ではなく「1億30万」と表示する)という慣例があるが、このゼロ非表示フラグ  $t$   $o$   $p$  はこの慣例に従った表示を実現するためのものである。また27は10進数を日本語の倍数表現を用いて表示した文字列としての日本語倍数表現配列  $w$   $h$   $e$   $r$   $e$   $[i]$  である。ここで、カウンタ  $i$  は日本語倍数表現を用いた文字列の先頭文字からの文字数の値である。

【0010】図3(a)は、本発明の一実施例における英語倍数表現パターンの構成図、図3(b)は本発明の一実施例における倍数表現英単語とその倍数との関係図であり、いずれも倍数表現記憶手段18に格納されている。図3(a)中、CARDは基数詞(例えば、「two」、「twenty-two」など、但し、倍数表現英単語を除く)、A\$は「hundred」、「thousand」、・・・などの倍数B(10のべき乗)を示す倍数表現英単語、ORDは「first」、「second」、「third」、・・・などの序数詞である。そして倍数表現英単語を用いた英語の数値表現は、図3(a)の各パターンのいずれかに該当する。

【0011】次に図4～図9を参照しながら、本実施例における形態素解析装置の処理の流れについて説明する。ここで図4～図8は、本発明の一実施例における形態素解析装置のフローチャート、図9(a)～(c)は本発明の一実施例における形態素解析装置の処理過程説明図である。まず図4は処理の概要を示している。

【0012】まずステップ1にて、英文文字列が入力され文字列記憶手段13に格納される。ここでは図9(a)に示すように「I saw twenty-two hundred and five peoples in the hall.」という英文文字列が入力されたものとする。次に、トークン分割手段14がこの英文字

## 4

列をトークン毎に分割したトークンリストを作成しトークン記憶手段15に格納する(ステップ2)。ここでは、トークンリストは図9(b)に示すように、トークン毎に前後双方向のポインタ(矢印)で連結した構成となっており、先頭のトークン「I」と末尾のトークン「。」の一方のポインタは便宜上「NULL」を指すこととしている。

【0013】次にステップ3にて各トークンについて辞書情報が獲得され、ステップ4にて図5に示す倍数表現処理(詳細は後述)が行われる。その結果、英語倍数表現が英文文字列中に存在すれば10進数生成手段21が10進数を生成する。そこで処理後の英文文字列について10進数が存在するかどうか調べ(ステップ5)、存在すれば、10進数を図6～図8に示す日本語倍数表現処理(詳細は後述)を施し(ステップ6)、以上を終了までくり返す(ステップ7)。

【0014】次にステップ4の倍数表現処理について図5に基づき説明する。ここで図5では倍数表現英単語A\$として「hundred」を採用した処理を示しているが、他の倍数表現英単語についても同様の処理により対応できる。

【0015】さてステップ11にて、英文文字列に「hundred」が存在するかどうか調べる。なければ「hundred」についての処理を終える。一方、「hundred」が存在すれば、「hundred」の前にCARD(基数詞)があるかどうか調べる(ステップ12)。なければ「hundred」をそのまま「ひゃく(百)」又は「100」と訳せばよいから「hundred」の処理を終え、あれば「hundred」の前のCARDを変数  $s$   $t$   $a$  に格納する(ステップ13)。また一旦「hundred」を変数  $e$   $n$   $d$  に格納する(ステップ14)。次にステップ15にて、「hundred」の次がCARDかどうか調べる。CARDであれば「hundred」の次のCARDを変数  $e$   $n$   $d$  に格納し(ステップ16)、ステップ17へ進む。この場合、図3(a)のNo. 3のパターンとなる。一方、ステップ15の判定が「否」であれば、「hundred」の次が「and」であるかどうか調べる(ステップ18)。「and」でなければステップ17へ進む。「and」であれば「and」の次がCARDかどうか調べる(ステップ19)。そうであれば「and」の次のCARDを変数  $e$   $n$   $d$  に格納する(ステップ20)。そしてステップ17にて、変数  $s$   $t$   $a$  から変数  $e$   $n$   $d$  までを、10進数生成手段21が10進数に変換する。この変換は、「and」を和記号(+)におきかえるとともに、 $CARD \times (A\$ \text{ に対応する倍数 } B)$  という演算により行う。

【0016】ここで図9(b)のトークンリストの例では、「hundred」が存在し(ステップ11)、「hundred」の前は「twenty-two」と

## 5

いうCARDであるから(ステップ12)、変数staに「twenty-two」というCARDが格納される(ステップ13)。そして一旦「hundred」が変数endに格納され(ステップ14)、「hundred」の次はCARDでなく"and"であるから(ステップ15, 18)、「and」の次の「five」というCARDが変数endに格納される(ステップ19, 20)。そしてステップ17の10進数変換が行われるのであるが、この変換により、変数staから変数endまでの「twenty-two hundred and five」という表記が「22×100+5」=「2205」という10進数の文字列に置き換えられる。

【0017】次にステップ6の10進数を日本語の倍数表現に変更する日本語倍数表現処理について、図6～図8を参照しながら説明する。図6では「億」以上の桁の処理を示す。まずステップ30にて、日本語倍数表現生成手段22は10進数を入力する。次にステップ31にて、日本語倍数表現配列where[i] (以下単に配列where[i]という)のカウンタiに0をセットし、倍数表現付加フラグflg (以下単にフラグflgという)に0 (付加しない)をセットし、注目桁変数C (以下単に変数Cという)を先頭桁にし、残り文字列長変数len (以下単に変数lenという)を10進数の全桁数とする。そして、ステップ32にて変数lenが8を越えているか調べ、越えていれば1億以上の数値でありステップ33へ、越えなければステップ40へ移る。

【0018】ステップ33では、フラグflgに1をセットし、配列where[i]に変数Cを代入し、カウンタiをインクリメントする(ステップ33～35)。そして変数lenが4の倍数であれば配列where[i]にカンマ", "を代入する(ステップ36, 37)。以下この処理をカンマ処理という。ここで、カンマは4桁おきに付与されるので、このように変数lenが4の倍数かどうかチェックしている。次いでステップ34にて10進数の1桁分の処理がすんだので、変数lenをデクリメントし(ステップ38)、注目する桁を1桁くり下げ(ステップ39)、ステップ32に戻る。

【0019】一方ステップ40では、変数flgが1かどうか調べ、1でなければ図7の処理へ移る。1であれば、億以上の桁が存在したため、ステップ33においてフラグflgに1がセットされたものであり、配列where[i]の現在の処理中の桁に「億」をセットし(ステップ44)、カウンタiをインクリメントして(ステップ42)、図7の処理へ移る。

【0020】図7のステップ43では、フラグtopを1 (ゼロ非表示)とし、ステップ44ではフラグflgを0 (倍数表現を付加しない)としてステップ45へ移る。ステップ45では変数lenが4を越えているかど

## 6

うか調べる。越えているならば、残りの桁が万以上であることになる。そしてステップ46では、フラグtopが1でありかつ変数Cが0であるかどうか調べる。これが是であれば、変数lenをデクリメントし(ステップ47)、変数Cを1桁下げて(ステップ48)ステップ45へ戻る。この処理により、「億」に直接後続する0を省略することができる。

【0021】ステップ46の判断が否であれば、フラグtopを0 (ゼロ表示)とし(ステップ52)、フラグflgに1をセットし(ステップ53)、配列where[i]の現在注目している桁に変数Cをセットし(ステップ54)、カウンタiをインクリメントする(ステップ55)。そしてステップ56～58において前述したカンマ処理(ステップ36～37と同様)を行ってステップ47へ戻る。一方、ステップ45において変数lenが4を越えないときは、残りの桁が万を越えていないものであり、ステップ49にてフラグflgが1 (倍数表現を付加する)ならば、配列where[i]の現在注目している桁に「万」をセットし(ステップ50)、カウンタiをインクリメントして(ステップ51)、図8の処理へ移る。一方、ステップ59にてフラグflgが1ならばそのまま図8の処理へ移る。次に図8では千以下の桁について図7と同様の処理が行われる(ステップ59～69)。以上の処理が、変数lenが0になるまで行われたら図4のステップ7へ戻る。

【0022】なお、本実施例では、原文を英語として本発明を説明したが、英語と同様な文法形態を有する外国語であれば本発明の効果が得られることは言うまでもない。

## 【0023】

【発明の効果】本発明の形態素解析装置は、原文文字列を入力する入力手段と、入力手段から入力された原文文字列を記憶する文字列記憶手段と、文字列記憶手段に記憶された原文文字列をトークン毎に分割するトークン分割手段と、トークン分割手段により分割されたトークンを記憶するトークン記憶手段とを備え、原語倍数表現を記憶する倍数表現記憶手段と、トークン記憶手段に記憶されたトークンと倍数表現記憶手段に記憶された原語倍数表現とを参照して、数値を表現するトークンから、そのトークンに相当する10進数を生成する10進数生成手段とを有するので、原語特有の倍数表現を10進数に変換し、正しく形態素解析することができる。

## 【図面の簡単な説明】

【図1】本発明の一実施例における形態素解析装置のブロック図

【図2】本発明の一実施例における形態素解析装置の機能ブロック図

【図3】(a) 本発明の一実施例における英語倍数表現パターンの構成図

(b) 本発明の一実施例における倍数表現英単語とその

## 倍数との関係図

【図 4】本発明の一実施例における形態素解析装置のフローチャート

【図 5】本発明の一実施例における形態素解析装置のフローチャート

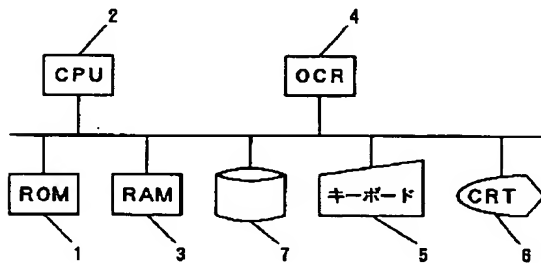
【図 6】本発明の一実施例における形態素解析装置のフローチャート

【図 7】本発明の一実施例における形態素解析装置のフローチャート

【図 8】本発明の一実施例における形態素解析装置のフローチャート

【図 9】(a) 本発明の一実施例における形態素解析装

【図 1】



【図 3】

(a)

No.	英語倍数表現パターン
1	CARD A\$
2	CARD A\$+"th"
3	CARD A\$ CARD
4	CARD A\$ "and" CARD
5	CARD A\$ "and" ORD

(b)

No.	倍数表現英単語 A\$	倍数 B
1	hundred	100
2	thousand	1000
3	million	1000000
.	.	.
N	.	.

## 置の処理過程説明図

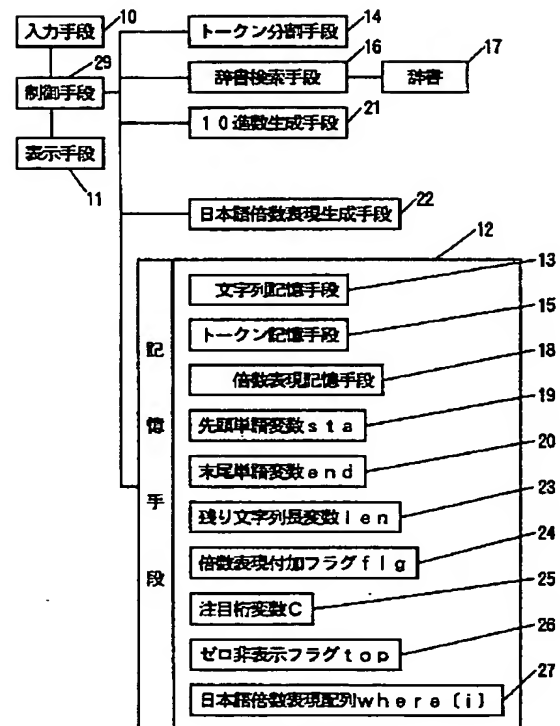
(b) 本発明の一実施例における形態素解析装置の処理過程説明図

(c) 本発明の一実施例における形態素解析装置の処理過程説明図

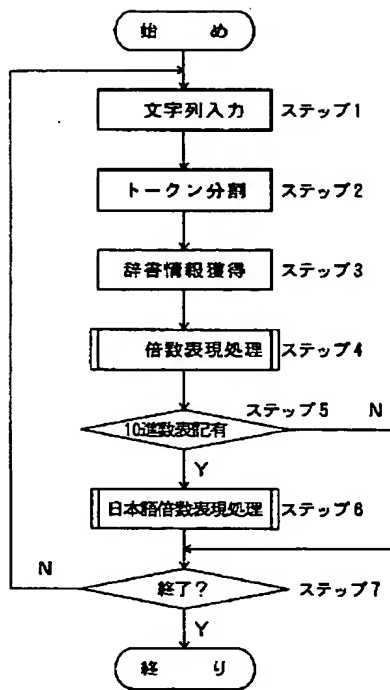
## 【符号の説明】

- 10 入力手段
- 13 文字列記憶手段
- 14 トークン分割手段
- 15 トークン記憶手段
- 18 倍数表現記憶手段
- 21 10進数生成手段

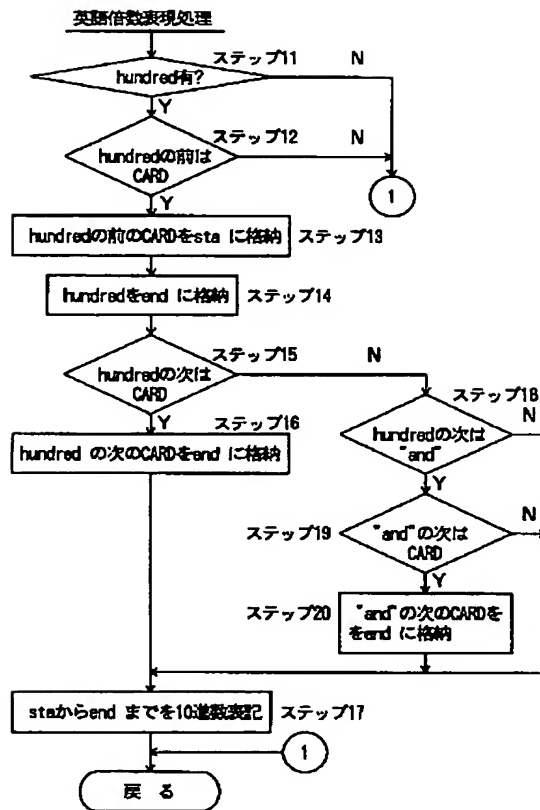
【図 2】



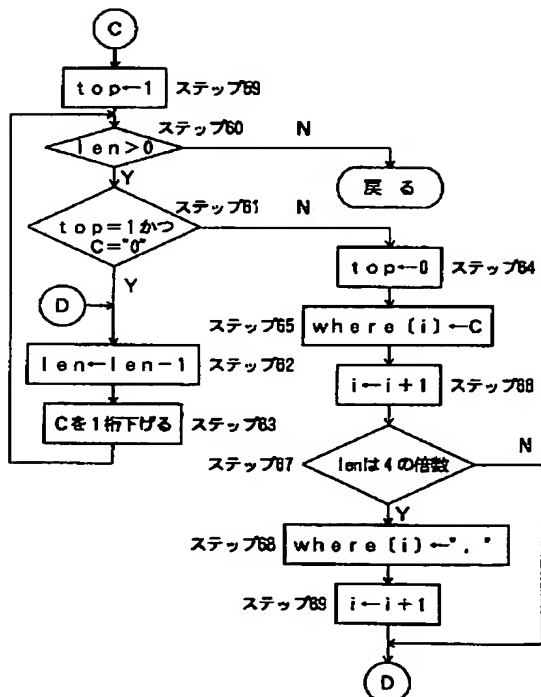
【図4】



【図5】

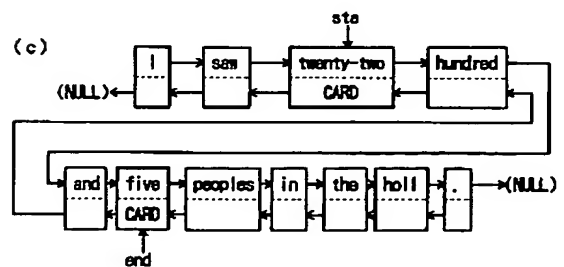
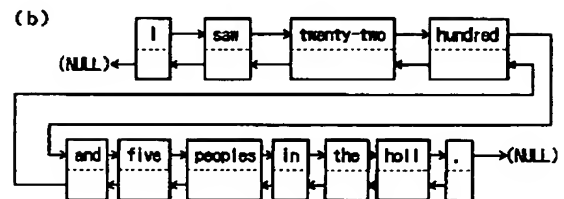


【図8】

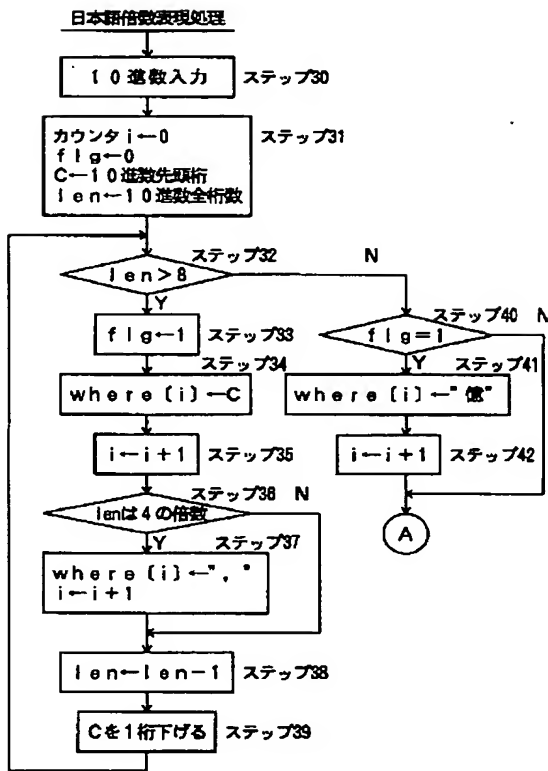


【図9】

(a) I saw twenty-two hundred and five peoples in the hall.



【図6】



【図7】

